WLAM- Attention: Plug-and-Play Wavelet Transform Linear Attention

Bo Feng , Chao Xu * , Zhengping Li , Shaohua Liu

Posted Date: 26 February 2025

doi: 10.20944/preprints202502.2130.v1

Keywords: Vision Transformer; Wavelet Transform; Self-attention Learning; Image Recognition

Article

WLAM- Attention: Plug-and-Play Wavelet Transform Linear Attention

Bo Feng [1,2], Chao Xu [1,2,*] , Zhengping Li [1,2] and Shaohua Liu [3]

[1] School of Integrated Ciruits, Anhui University, Hefei 230601, China; p18101008@stu.ahu.edu.cn (B.F.);04173@ahu.edu.cn (Z.L.)

[2] Anhui Engineering Laboratory of Agro-Ecological Big Data, Hefei 230601, China

[3] Anhui Zhongke Jingle Technology Co., Ltd,Hefei 230601, China;shliu@zkjg.com(S.L.)

* Correspondence: xchao@ahu.edu.cn; Tel.: +86-133-3919-9368

Abstract: Linear attention has gained popularity in recent years due to its lower computational complexity compared to SoftMax attention. However, its relatively lower performance has limited its widespread application. To address this issue, we propose a plug-and-play module called Wavelet-Enhanced Linear Attention Mechanism (WLAM), which integrates Discrete Wavelet Transform (DWT) with linear attention. This approach enhances the model's ability to express global contextual information while improving the capture of local features.Firstly, we introduce DWT into the attention mechanism to decompose the input features. The original input features are utilized to generate the query vector Q, while the low-frequency coefficients are used to generate the key K. The high-frequency coefficients undergo convolution to produce the value V. This method effectively embeds global and local information into different components of the attention mechanism, thereby enhancing the model's perception of details and overall structure.Secondly, we perform multi-scale convolution on the high-frequency wavelet coefficients and incorporate a Squeeze-and-Excitation (SE) module to enhance feature selectivity. Subsequently, we utilize the Inverse Discrete Wavelet Transform (IDWT) to reintegrate the multi-scale processed information back into the spatial domain, addressing the limitations of linear attention in handling multi-scale and local information.Finally, inspired by certain structures of the Mamba network, we introduce a forget gate and an improved block design into the linear attention framework, inheriting the core advantages of the Mamba architecture. Following a similar rationale, we leverage the lossless downsampling property of wavelet transforms to combine the downsampling module with the attention module, resulting in the Wavelet Downsampling Attention (WDSA) module. This integration reduces the network size and computational load while mitigating information loss associated with downsampling.We apply Wavelet-Enhanced Linear Attention Mechanism (WLAM) to classical networks such as PVT, Swin, and CSwin, achieving significant improvements in performance on image classification tasks. Furthermore, we combine Wavelet Linear Attention with the Wavelet Downsampling Attention (WDSA) module to construct WDLMFormer, which achieves an accuracy of 84.2% on the ImageNet-1K dataset.

Keywords: Vision Transformer; Wavelet Transform; Self-attention Learning; Image Recognition.

1. Introduction

In recent years, Transformer models have shown remarkable performance in the field of computer vision, achieving significant success in image classification, object detection, semantic segmentation, and multimodal tasks. However, the use of Transformers and self-attention mechanisms in computer vision still faces considerable challenges.Modern Transformer models typically employ the Softmax attention mechanism, which calculates the similarity between each query-key pair. The computational complexity of this mechanism grows quadratically with the number of tokens. As a result, the Softmax attention mechanism can lead to uncontrollable

 doi:10.20944/preprints202502.2130.v1

computational demands.The self-attention mechanism lacks the inductive biases found in CNNs, such as translation invariance and locality. These inductive biases are crucial for the model's generalization ability on smaller datasets, and the absence of such features in Transformers may affect their performance on certain tasks.

To address the uncontrollable computational demands posed by the Softmax attention mechanism, various remedial measures have been proposed in prior work. PVT[2] introduced sparse global attention by reducing the resolution of keys (K) and values (V) to manage computational costs. The Swin-Transformer[3] alleviated the computational burden by limiting self-attention calculations to local windows, thereby reducing the receptive field. Subsequently, Swin-Transformer_V2[4] improved accuracy under large sample conditions. DAT[5] proposed a deformable attention mechanism that adaptively focuses on different regions of the input features. NAT[6] simulated convolutional operations and presented an automated network design approach based on the Transformer architecture. BiFormer[7] employed dual-level routing attention to dynamically identify areas of interest for each query.However, these methods inherently restrict the overall receptive field of self-attention or are heavily influenced by specially designed attention patterns, hindering their plug-and-play adaptability. Linformer[8] discarded the Softmax function and decoupled it into two independent function ϕ, allowing the attention computation order to shift from (query · key) · value to query · (key · value), thereby reducing the overall computational complexity to O(N). Nevertheless, this approximation resulted in a significant performance drop[9,10].To mitigate this issue, Efficient Attention[11] proposed an effective attention mechanism that applies the Softmax function to both Q and K. SOFT[12] and Nyströmformer[13] further approximated the Softmax operation using matrix decomposition. Castling-ViT[14] utilized Softmax attention as a training auxiliary tool while exclusively employing linear attention during inference. FLatten-Transformer[15] introduced a focus function and leveraged deep convolutions to preserve feature diversity.Despite the effectiveness of these approaches, they still face limitations in expressive capacity due to the constraints of linear attention. Agent-Attention[16] defined a novel four-component attention mechanism (Q, A, K, V), where the agent vector A serves as a proxy for the query vector Q, aggregating information from K and V before broadcasting it back to Q. This agent-based attention mechanism enables the modeling of global information with significantly reduced computational costs.

To address the limitations of self-attention mechanisms in processing local information, various hybrid models combining Convolutional Neural Networks (CNNs) and Transformers have been proposed. Models such as SCTNet[17], AdaMCT[18], TransXNet[19], and Enriched CNN-Transformer[20] represent parallel fusion networks, where the architecture is divided into CNN and Transformer branches, and the information from both branches is integrated through a fusion network. In contrast, EdgeNeXt[21], CvT[22], MobileViT2[23], MobileViT3[24], and MLLA[25] are examples of serial fusion networks that first utilize CNNs to extract local features, which are then fed into a Transformer for global context modeling, thereby enhancing feature representation capabilities. However, these networks are specifically designed architectures, making them largely incompatible with plug-and-play applications.

This paper presents a plug-and-play WDTML_Attention mechanism, which leverages wavelet transform algorithms to enhance the global context representation capability of linear attention while maintaining effective local information expression.

Key Innovations:

- Integration of Discrete Wavelet Transform (DWT) with Linear Attention: The proposed method incorporates DWT into the attention mechanism by decomposing the input features for different attention constructs. Specifically, the input features are utilized to generate the attention queries (Q), low-frequency information is employed to generate the attention keys (K), and high-frequency information, processed through convolution, is used to generate the values (V). This approach effectively enhances the model's ability to capture both local and global features, improving the perception of details and overall structure;
- Multi-Scale Processing of Wavelet Coefficients: The high-frequency wavelet coefficients are processed through convolutional layers with varying kernel sizes to extract features at different

 doi:10.20944/preprints202502.2130.v1

scales. This is complemented by the Squeeze-and-Excitation (SE) module, which enhances the selectivity of the features. An inverse discrete wavelet transform (IDWT) is utilized to reintegrate the multi-scale decomposed information back into the spatial domain, compensating for the limitations of linear attention in handling multi-scale and local information;

- Structural Mimicry of Mamba Network: The proposed wavelet linear attention incorporates elements from the Mamba network, including a forget gate and a modified block design. This adaptation retains the core advantages of Mamba, making it more suitable for visual tasks compared to the original Mamba model;
- Wavelet Downsampling Attention (WDSA) Module: By exploiting the lossless downsampling property of wavelet transforms, we introduce the WDSA module, which combines downsampling and attention mechanisms. This integration reduces the network size and computational load while minimizing information loss caused by downsampling.

2. Related Work

Wavelet transform is an effective method for time-frequency analysis. It is reversible and capable of preserving a significant amount of information, making it widely applicable in various neural network architectures. For instance, Bae et al. [26] were among the first to incorporate wavelet transform into CNNs for image restoration tasks. In [27], Haar wavelets were integrated into CNNs for multi-resolution analysis, achieving texture classification and image labeling. Additionally, [28] introduced wavelet transform into Transformer models, demonstrating promising performance in image classification and object detection. This model reduced the number of input feature channels to one-fourth of the original, employing wavelet transform and convolution to generate the keys (K) and values (V) for Softmax attention, followed by a wavelet inverse transform to fuse the output features. However, this approach did not leverage the lossless downsampling properties of wavelet transforms to reduce computational complexity in the attention module, nor did it fully exploit the multi-resolution analysis capabilities of wavelet transforms.

In [29], wavelet transform was utilized for downsampling at the front end of the model, with inverse wavelet transform used for upsampling at the end. This method effectively lowered the image resolution while preserving significant image features, leading to reduced resource consumption in Transformer models. Yet, it inadequately utilized the multi-resolution analysis potential of wavelet transforms. The work in [30] made minor modifications to the approach in [28] and applied them within a U-Net architecture, but it suffered from the same limitations. In [31], a multi-scale enhancement module was developed using wavelet transform, convolution, nonlinear transformations, and inverse wavelet transform to enhance the multi-scale recognition capabilities of neural networks. Furthermore, [32] employed gradient wavelet transform and Transformer networks to improve edge information recognition. Lastly, [39] proposed a novel wavelet-based Mamba model with Fourier adjustment, termed WalMaFa, which consists of a wavelet-based Mamba block (WMB) and a fast Fourier adjustment block (FFAB), achieving outstanding initial brightness enhancement.

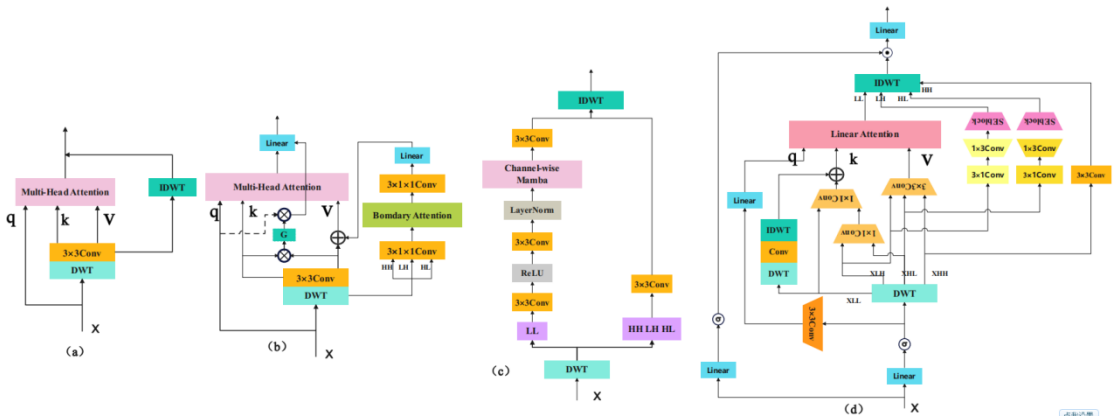


Figure 1. Wavelet Transform Enhanced Transformer Structure Diagram: (a) Attention Diagram from [28]; (b) Attention Diagram from [30]; (c) Attention Diagram from [42]; (d) Attention Diagram from this paper.

 doi:10.20944/preprints202502.2130.v1

3. Our Work

3.1. Plug-and-Play WLAM Attention Module

As indicated in [29], wavelet transform can achieve nearly lossless downsampling, thereby reducing the computational complexity of neural networks. Additionally, insights from [31] and [32] demonstrate that utilizing the multi-resolution analysis capabilities of wavelet transforms can significantly enhance a neural network's ability to recognize local details and edge features. The WLAM (Wavelet-Enhanced Linear Attention Mechanism) designed in this paper fully exploits the lossless downsampling and multi-resolution analysis capabilities of wavelet transforms, leading to a substantial increase in linear attention recognition capabilities while effectively reducing computational workload. This is particularly evident in the improvement of the module's ability to express local information.

In both [28] and [30], the integration of Softmax Attention with wavelet transform is employed to lower computational complexity and achieve multi-resolution analysis. They utilize the input features X as the Query (Q), compressing the number of channels through a linear transformation to one-fourth of the original, and subsequently obtaining the Key (K) and Value (V) through wavelet transform and convolution, as illustrated in the figure. However, Softmax Attention employs exponentially weighted normalization for the calculation of attention weights, which is computed as follows:

$$\mathrm{Softmax}\left(\frac{(QK^T)}{\sqrt{d_k}}\right)_{ij} = \frac{\exp\left(\frac{(QK^T)_{ij}}{\sqrt{d_k}}\right)}{\sum_{k=1}^{m}\exp\left(\frac{(QK^T)_{ij}}{\sqrt{d_k}}\right)} \tag{1}$$

In this context, exp() denotes the exponential function, $\sum_{k=1}^{m}\exp()$ representing the sum of all exponential functions. This normalization process tends to significantly amplify features with higher weights while rendering those with lower weights nearly negligible. Consequently, Softmax Attention is relatively sensitive to the feature distributions of Q and K, necessitating that they reside in similar spaces. As illustrated in Figures 1(a) and 1(b) from [28] and [30], the feature distributions of Q and K can exhibit significant disparities, which prevents them from existing in a comparable space. This discrepancy can lead to computational instability.

From equation (2), we observe that the wavelet transform decomposes the input tensor X into four sub-bands, resulting in both the height (H) and width (W) being reduced to half of their original dimensions, specifically H/2 and W/2.

$$\{X_{LL}, X_{LH}, X_{HL}, X_{HH}\} = \mathrm{DWT}(X) \tag{2}$$

- X_{LL}:This sub-band preserves most of the image's energy and structural information, making it the richest in content. Typically, the primary structures and general shapes of the image are contained within this sub-band.
- X_{LH}:This sub-band represents the horizontal details of the image, capturing high-frequency components such as horizontal edges or textures. However, it contains relatively less information, primarily focusing on changes in the horizontal direction.
- X_{HL}:This sub-band captures the vertical details of the image, including vertical edges and textures. However, it contains relatively less information, primarily focusing on variations in the vertical direction.
- X_{HH}:This sub-band represents the finest details of the image, encompassing diagonal features such as diagonal edges. It contains high-frequency noise and very subtle details, resulting in the least amount of information.

The information carried by the four sub-bands resulting from wavelet decomposition is unevenly distributed, with the majority of the information concentrated in the X_{LL} sub-band. In papers [28, 30], the authors compress the channels of the input X to D/4 and then expand the channel count back to D through wavelet transformation, ultimately obtaining K and V through a convolution. Due to the uneven distribution of information across the sub-bands, a significant

 doi:10.20944/preprints202502.2130.v1

amount of information is lost during the process of first compressing the channel count and then expanding it through wavelet transformation. As a result, the information content in K and V is considerably lower than that in Q. Therefore, while the method in papers [28, 30] appears to reduce computational complexity and enhance sensitivity to local information by leveraging wavelet transforms, it does not fully utilize the advantages of the wavelet transform.

To maximize the potential of the multi-resolution analysis afforded by wavelet transforms, we have opted to forgo the Softmax Attention, which is relatively sensitive to the feature distribution of Q and K, and instead employ a linear attention mechanism combined with wavelet transformation. The key characteristic of linear attention is that it typically utilizes the dot product of Q and K directly, along with kernelization, without applying **Softmax** normalization.

$$Q = \phi(xW_q), K = \phi(xW_k), V = \phi\emptyset(xW_v) \tag{3}$$

$$\text{linear\_Attention} = \sum_{i=1}^{N} \frac{Q_i K_j^T}{\sum_{j=1}^{N} Q_i K_j^T} V_j = \frac{Q_i(\sum_{j=1}^{N} K_j^T V_j)}{Q_i(\sum_{j=1}^{N} K_j^T)} \tag{4}$$

From equation (4), it can be observed that the linear attention weights are accumulated linearly, rather than amplifying or diminishing the weights of specific features. In other words, the weighting mechanism of linear attention is more balanced, making it suitable for handling more diverse inputs. By applying the associative property of matrix multiplication, we can rearrange the computation order from the Softmax Attention format $(QK^T)V$ to $Q(K^TV)$, thereby reducing the computational complexity to O(N). This represents a significant decrease in computational complexity compared to Softmax Attention.

In traditional self-attention mechanisms, the input X is typically subjected to linear transformations to generate queries (Q), keys (K), and values (V), as shown in equation (3). In contrast, we utilize wavelet transformation to achieve a more diverse input representation.

We downsample X to serve as the query(Q), defined as $Q_{dwt} = \phi(Conv_{3\times3}(X)_{\downarrow 2}W^q)$. The original feature map X contains all the raw information and is typically well-suited as a query, as queries are designed to capture the global information of the input data for calculating attention weights.

In contrast to the approaches presented in papers [28, 30], we first apply wavelet transformation to the input feature X as $\{X_{LL}, X_{LH}, X_{HL}, X_{HH}\} = DWT(X)$, and then we proportionally compress the information contained in each sub-band to obtain the keys (K).

$$K = Conv_{1\times1}^{1.5C\to C}(\,Concat(X_{LL}, Conv_{1\times1}^{C\to 0.25C}(X_{LH}), Conv_{1\times1}^{C\to 0.25C}(X_{HL})) \tag{5}$$

This approach allows K to provide a rich representation of features while minimizing information loss, which aids the model in capturing more useful information within the attention mechanism. Additionally, we can perform a secondary wavelet transformation here to further enhance the low-frequency sub-band X_{LL}, thereby mimicking the scale enhancement module presented in paper [31].

$$\{DX_{LL}, DX_{LH}, DX_{HL}, DX_{HH}\} = DWT(X_{LL}) \tag{6}$$

$$DX_{LL1} = Conv_{7\times7}(DX_{LL}) \tag{7}$$

$$DX_{LH1} = Conv_{3\times1}(Conv_{1\times3}(DX_{LH})) \tag{8}$$

$$DX_{HL1} = Conv_{3\times1}(Conv_{1\times3}(DX_{HL})) \tag{9}$$

$$K_{dwt} = \phi((IDWT\{DX_{LL1}, DX_{LH1}, DX_{HL1}, DX_{HH}\}+K)W_k) \tag{10}$$

We use the high-frequency sub-bands X_{LH}, X_{HL}, and X_{HH} to create V (Value) through convolution, defined as $V_{dwt} = \phi(Conv_{3\times3}^{3C\to C}(Concat(X_{LH}, X_{HL}, X_{HH})))W^V$. Here, X_{LH}, X_{HL}, and X_{HH} represent the high-frequency coefficients extracted from the wavelet decomposition of the feature map X, with X_{LH} corresponding to horizontal high-frequency coefficients, X_{HL} to vertical high-

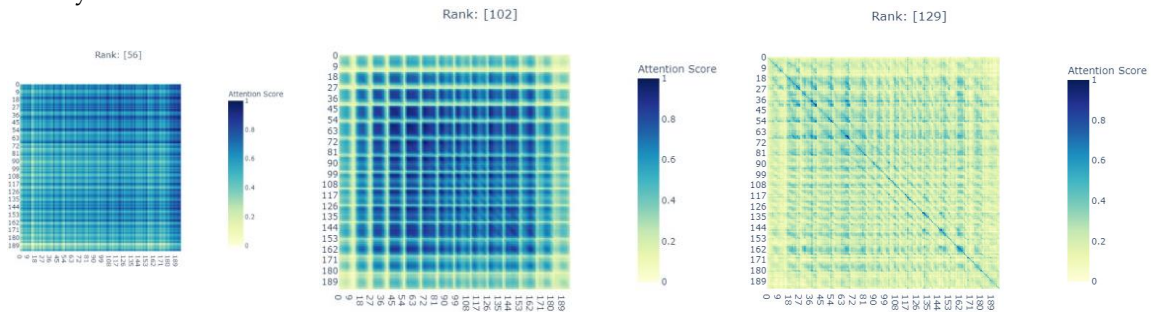
 doi:10.20944/preprints202502.2130.v1

frequency coefficients, and X_{HH} to diagonal high-frequency coefficients. The improved V consists entirely of high-frequency information, excluding low-frequency components, which enhances the attention module's ability to capture local features more effectively.

$$\text{linear\_Attention}_{dwt} = Q_{dwt}(K_{dwt}^{T}V_{dwt}) \tag{11}$$

By employing the aforementioned approach, we provide multi-resolution representations for Q, K, and V in the linear attention mechanism, thereby enhancing the model's performance when dealing with diverse and complex inputs. Let the input tensor X have dimensions H and W; consequently, the dimensions of Q, K, and V will be H/2 and W/2. This further reduces the computational burden of linear attention.

The improvements outlined above can enhance the expressive capability of linear attention to some extent; however, linear attention still exhibits suboptimal performance in terms of feature diversity.



(**a**) (**b**) (**c**)

Figure 2. illustrates the rank of the attention matrices: (**a**) rank of the linear attention matrix; **(b)** rank of the attention matrix after the addition of a Depthwise Convolution (DWC) module; and **(c)** rank of the attention matrix following the inverse wavelet transformation.

Rank of the attention matrix: In traditional Transformer models, the attention matrix is typically of full rank, indicating a high degree of feature diversity.

$$\text{rank}(\text{Softmax}(QK^{T})) = N \tag{12}$$

The rank in linear attention is constrained by the number of tokens **N** in each head and the channel dimension **d**, as illustrated in Figure 2(a).

$$\text{rank}(QK)^{T}) \leq \min\{\text{rank}(Q), \text{rank}(K)\} \leq \min\{N, d\} \tag{13}$$

Since d is typically less than N, the rank of the attention matrix in the linear attention mechanism is limited to d, whereas the softmax attention can be ranked up to N (and is likely to be equal to both d and N). In this context, the upper bound of the attention matrix's rank is constrained to a lower ratio, indicating that many rows of the attention mapping are severely homogenized. As the output of self-attention is a weighted sum of the same set of V, the uniformity of attention weights inevitably leads to similarities among the aggregated features.

To address this issue, papers [15] and [16] propose the incorporation of a Depthwise Convolution (DWC) module in the attention matrix, with the output represented as follows:

$$\text{Out} = QK^{T}V + \text{DWC}(V) = (QK^{T} + M_{\text{DWC}})V \tag{14}$$

$$\text{DWC}(V) = \text{Conv}_{3\times3}(V) \tag{15}$$

Here, M_{DWC} is a sparse matrix corresponding to the depthwise convolution operation. Since M_{DWC} has the potential to become a full-rank matrix, it effectively raises the upper limit of the rank of the equivalent attention. As shown in Figure 2(b), although this approach results in a significant increase in the rank value, the actual improvement in model accuracy is quite limited.

This paper proposes enhancing the high-frequency components $X_{LH}, X_{HL}, \text{and } X_{HH}$ using a depthwise convolution module, followed by the integration of these enhancements with Linear Attention through inverse wavelet transformation. This method significantly improves the feature

 doi:10.20944/preprints202502.2130.v1

diversity of Linear Attention. In contrast to the direct addition of a depthwise convolution module as suggested in papers [15] and [16], the inverse wavelet transformation enables a more effective fusion of the features from Linear Attention and the depthwise convolution module.The components X_{LH}, X_{HL}, and X_{HH} capture most of the local features present in the input tensor X. By enhancing these components with a depthwise convolution module, we can substantially improve the module's capability to extract local features. Paper [35] highlights that performing convolution in the wavelet domain results in a larger receptive field. By combining inverse wavelet transformation with Linear Attention, we can significantly address the deficiencies in Linear Attention's ability to extract local information, thereby achieving robust extraction capabilities for both global and local features.

$$X_{LH1} = \mathrm{Relu}(\mathrm{Conv}_{1\times1}^{2C\to C}(\mathrm{SE}(\mathrm{Conv}_{1\times3}^{C\to 2C}(\mathrm{Conv}_{3\times1}(X_{LH})))) + \mathrm{BN}(\mathrm{Conv}_{1\times1}(X_{LH})) \quad (16)$$

$$X_{HL1} = \mathrm{Relu}(\ \mathrm{Conv}_{1\times1}^{2C\to C}(\mathrm{SE}(\mathrm{Conv}_{1\times3}^{C\to 2C}(\mathrm{Conv}_{3\times1}(X_{HL})))) + \mathrm{BN}(\mathrm{Conv}_{1\times1}(X_{LH}))) \quad (17)$$

$$\{X_{LL}, X_{LH}, X_{HL}, X_{HH}\} = \mathrm{DWT}(X) \quad (18)$$

$$O_{idwt} = \mathrm{IDWT}(\mathrm{linear\_Attention}_{dwt}, (X_{LH1}, X_{HL1}, X_{HH1})) \quad (19)$$

Paper [42] shares a similar approach to ours; however, its method employs only a single 3×3 convolution across all high-frequency subbands, which limits the effective extraction of local features within those subbands. Drawing inspiration from the architecture of MobileNetV3 [33], we note that X_{LH}contains only horizontal local features while X_{HL} contains only vertical local features. Therefore, we utilize both $\mathrm{Conv}_{3\times1}$ and $\mathrm{Conv}_{1\times3}$ to extract features while reducing computational complexity. The channel dimension is expanded to 2C, followed by the application of a Squeeze-and-Excitation (SE) attention module, after which a $\mathrm{Conv}_{1\times1}$ is used to reduce the channel count back to C. This approach substantially enhances the model's nonlinear expressive capability.

The X_{HH} subband contains high-frequency noise and very fine details, representing the least amount of information, which is why we apply only a 3×3 convolution to it. Ultimately, we treat Linear Attention as the low-frequency subband while combining X_{LH}, X_{HL}, and X_{HH} as high-frequency subbands to perform the inverse wavelet transformation, resulting in the output O_{idwt}. As illustrated in Figure 2(a), the rank of the attention module significantly increases after the inverse wavelet transformation, with the dimensions of the output feature tensor restored to H and W.

Furthermore, paper [34] introduces Mamba, which can be viewed as a variant of the linear attention Transformer characterized by specialized linear attention and an improved block design. By integrating certain structural elements of Mamba into our linear attention framework, we can enhance its performance. Building on this foundation, we incorporate the superior architecture of Mamba [36] into our WLAM attention module.

1. Drawing inspiration from the forget gate in Mamba, we modified the query Q_{dwt} and key K_{dwt} mechanisms. The forget gate imparts two essential attributes to the model: local bias and positional information. In this study, we replaced the forget gate with Repositioned Positional Encoding (RopE), which yielded an accuracy improvement of 0.5%;

$$Q_{dwt} = \mathrm{RopE}(Q_{dwt}) \quad (20)$$

$$K_{dwt} = \mathrm{RopE}(K_{dwt}) \quad (21)$$

2. Inspired by Mamba, we incorporated a learnable shortcut mechanism into the linear attention framework. This enhancement resulted in an accuracy improvement of 0.2%.

$$OUT_{TWMA} = \mathrm{liner}(O_{idwt} \cdot \delta(\ \mathrm{liner}(X))) \quad (22)$$

The proposed WLAM-Attention module significantly reduces the computational burden of the linear attention mechanism while enhancing input diversity through wavelet transformation. By employing deep convolution on the high-frequency subbands and utilizing inverse wavelet transformation, we effectively address the issue of limited feature diversity in linear attention. The WLAM-Attention module can be integrated as a plug-and-play component and is easily adaptable to various modern Vision Transformer (ViT) architectures. To demonstrate its effectiveness, the

 doi:10.20944/preprints202502.2130.v1

authors empirically applied the WLAM-Attention module to four advanced and representative transformer models, including PVT [2], Swin [3], and CSWin [41]. Detailed structural information can be found in Appendix A.

3.2. Lossless Downsampling Attention Module

The primary advantage of wavelet transformation lies in its ability to perform nearly lossless downsampling. Leveraging this property, we propose merging the downsampling module with the first attention module that follows the downsampling process into a single attention module, termed the Wavelet Downsampling Attention Module. This integration reduces computational complexity while minimizing information loss associated with downsampling. Let X denote the input tensor, with C representing the number of channels and H and W denoting the height and width, respectively.

$$X = \delta(\text{linear}(X)) \tag{23}$$

$$Q = \phi(\text{Conv}_{3\times3}^{C\to2C}(X)W^{q}) \tag{24}$$

$$\{X_{LL}, X_{LH}, X_{HL}, X_{HH}\} = \text{DWT}(X) \tag{25}$$

$$K = \phi((\text{Concat}(X_{LL}, \text{Conv}_{1\times1}^{C\to0.5C}(X_{LH}), \text{Conv}_{1\times1}^{C\to0.5C}(X_{HL}))W^{K}) \tag{26}$$

$$V = \phi(\text{Conv}_{3\times3}^{3C\to2C}(\text{Concat}(X_{LH}, X_{HL}, X_{HH})))W^{V} \tag{27}$$

$$\text{Wavelet\_Downsampling\_Attention} = Q(K^{T}V) + \text{DWC}(V) \tag{28}$$

The Wavelet Downsampling Attention module has a channel count of 2C and reduces the dimensions to H/2 and W/2.

3.3. Macro Architecture Design

The Wavelet-Enhanced Linear Attention Mechanism (WLAM) can be integrated as a plug-in component within various modern Vision Transformer (ViT) architectures, or it can be combined with the Sampling Attention Module to form the WLAMFormer network, as illustrated in Figure 3. The input is a natural image with dimensions $H \times W \times 3$. The image undergoes downsampling through a convolutional layer with a stride of 2, followed by another convolutional layer with a stride of 1, resulting in a downsampled output of size $\frac{H}{2} \times \frac{W}{2} \times C_0$, where C_0 represents the number of channels. Subsequently, the image is processed through four stages of encoding layers, with each stage utilizing ownSampling downsampling to produce feature maps of sizes $\frac{H}{4} \times \frac{W}{4} \times C_1$, $\frac{H}{8} \times \frac{W}{8} \times C_2$, $\frac{H}{16} \times \frac{W}{16} \times C_3$ and $\frac{H}{32} \times \frac{W}{32} \times C_4$, where C_i denotes the channel count for each feature map. Each stage consists of N_i stacked blocks, as depicted in the Figure3. The design is inspired by EfficientViT [37] and EdgeViT [38] networks, incorporating both the Wavelet Linear Attention module and the MLP module. For specific parameter settings, please refer to Appendix B and Figure 3.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 February 2025 doi:10.20944/preprints202502.2130.v1

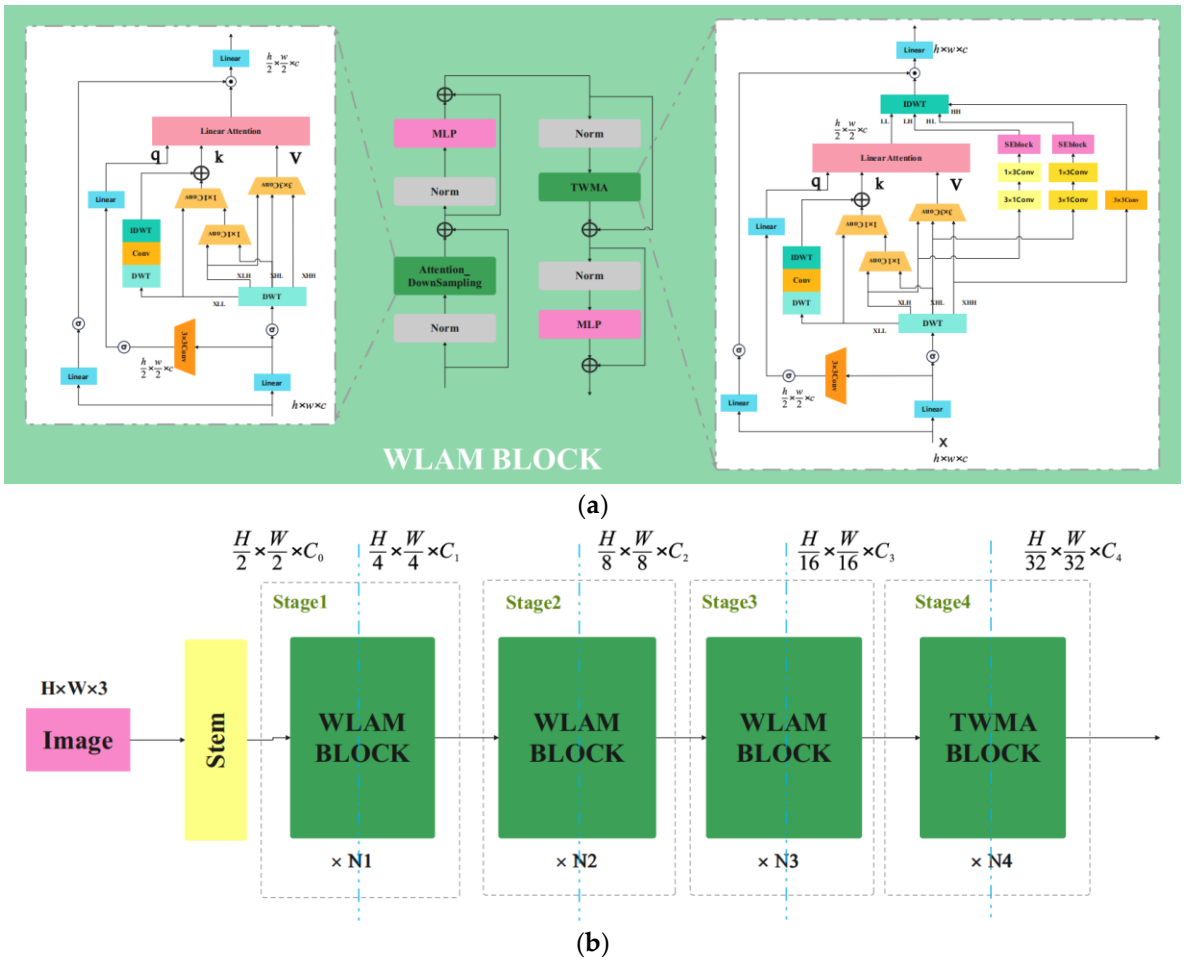


Figure 3. Illustrates the macro architecture design of the WLAMFormer network: (a) a schematic diagram of the TWMA BLOCK structure; (b) an overview of the overall structure of the WLAMFormer.

When the input image size is 224×224, we have $\frac{H}{32} = \frac{W}{32} = 7$. Due to the constraints of wavelet transformation, there is a minimum size requirement for the input image, which prevents the use of the WLAM attention module in stage 4. Consequently, we substitute it with a Linear Attention module.

4. Experiments

4.1. Image Classification

The ImageNet-1K dataset [40] comprises over 1.3 million images spanning 1,000 natural categories. Due to its diversity, this dataset covers a wide range of objects and scenes, making it one of the most widely used datasets in the field. We trained our network from scratch without utilizing any additional data, employing the CSwin-B model [41], which is pre-trained on ImageNet and achieves a top-1 accuracy of 84.2%, as the teacher model for distillation.

The training strategy follows the setup outlined in EdgeNeXt [21]. All models were trained with an input size of 224×224 using the AdamW [42] optimizer for 300 epochs, with a batch size of 1024. The learning rate was set to $\mathbf{1\times10^{-4}}$ with a cosine annealing schedule [43], and a warm-up period of 20 epochs was implemented. We enabled label smoothing (with a coefficient of 0.1) , random size cropping, horizontal flipping, RandAugment [44], and multi-scale sampling. During training, the exponential moving average (EMA) momentum was set to 0.9995. To fully leverage the network's effectiveness, we fine-tuned the model for an additional 30 epochs at a resolution of 384 ×384, using a learning rate of 1×10^{-5} and a batch size of 64.

We implemented the classification model based on PyTorch, running on six V100 GPUs. The experimental results on the ImageNet-1K dataset [40], presented in Table 1, clearly demonstrate the advancements our model brings to the field of image classification. It is important to note that for

 doi:10.20944/preprints202502.2130.v1

throughput, we report per-frame metrics on mobile devices and results with a batch size of 64 on GPUs. The results for all variants of our models are highlighted in bold.

Table 1. Comparison of Image Classification Performance on the ImageNet-1K Dataset.

Model	Par.↓(M)	Flops↓(G)	Throughput(A100)	Type	Top-1↑
PVTv2-B1[37]	**14.02**	**2.034**	**1945**	**Transformer**	**78.7**
SwiftFormer-L1[45]	12.05	1.604	**5051**	**Hybrid**	**80.9**
CAS_ViT_M[46]	**12.42**	**1.887**	**2254**	**Hybrid**	**82.8**
PoolFormer-S12[47]	**11.9**	**1.813**	**3327**	**Pool**	**77.2**
MobileViT-v2×1.5[23]	**10.0**	**3.151**	**2356**	**Hybrid**	**80.4**
EffiFormer-L1[11]	**12.28**	**1.310**	**5046**	**Hybrid**	**79.2**
WLAM**Former_L1**	**13.5**	**2.847**	**2296**	**DWT-Transformer**	**83.0**
ResNet-50	25.5	4.123	**4835**	**ConvNet**	**78.5**
PoolFormer-S24[47]	21.35	3.394	**2156**	**Pool**	**80.3**
PoolFormer-S36[47]	**32.80**	**4.620**	**1114**	**Pool**	**81.4**
SwiftFormer-L3[45]	**28.48**	**4.021**	**2896**	**Hybrid**	**83.0**
Swin-T[3]	28.27	4.372	**1246**	**Transformer**	**81.3**
PVT-S[2]	24.10	3.687	**1156**	**Transformer**	**79.8**
ConvNeXt-T[48]	**29.1**	**4.532**	**3235**	**ConvNet**	**82.1**
CAS-ViT-T[46]	21.76	3.597	**1084**	**Hybrid**	83.9
EffiFormer-L3[11]	**31.3**	**3.940**	**2691**	**Hybrid**	**82.4**
Vmanba-T[49]	**30.2**	**4.902**	**1686**	**Mamba**	**82.5**
MLLA-T[25]	**25.12**	**4.250**	**1009**	**mlla**	**83.5**
WTConvNeXt-T[50]	30M	4.5G	**2514**	**DWT-ConvNet**	82.5
WLAM**Former_L2**	**25.07**	**3.803**	**1280**	**DWT-Transformer**	**84.1**
ConvNeXt-S[48]	**50.2**	**8.74**	**1255**	**ConvNet**	**83.1**
PVTv2-B3[37]	**45.2**	**6.97**	**403**	**Transformer**	**83.2**
CSwin-S[41]	**35.4**	**6.93**	**625**	**Transformer**	**83.6**
VMamba-S[49]	**50.4**	**8.72**	**877**	**Mamba**	**83.6**
MLLA-S[25]	**47.6**	**8.13**	**851**	**mlla**	**84.4**
WTConvNeXt-S[50]	54.2	8.8G	**1045**	**DWT-ConvNet**	83.6
WLAM**Former_L3**	**46.6**	**7.75**	**861**	**DWT-Transformer**	**84.6**

Through Table 1 and Figure 4, we observe that the WLAMFormer model consistently achieves higher Top-1 accuracy compared to other models with similar computational budgets and parameter counts.

WLAMFormer_L1 (13.5M parameters) reaches a Top-1 accuracy of 83.0%, outperforming models such as CAS_ViT_M [46] (12.42M, 82.8%), SwiftFormer-L1 [45] (12.05M, 80.9%), and EffiFormer-L1 [11] (12.28M, 79.2%).

WLAMFormer_L2 (25.07M parameters) achieves a Top-1 accuracy of 84.1%, surpassing CAS-ViT-T [46] (21.76M, 83.9%), ConvNeXt-T [48] (29.1M, 82.1%), and Swin-T [3] (28.27M, 81.3%).

WLAMFormer_L3 (46.6M parameters) attains a Top-1 accuracy of 84.6%, exceeding MLLA-S [25] (47.6M, 84.4%), CSwin-S [41] (35.4M, 83.6%), and ConvNeXt-S [48] (50.2M, 83.1%).

These results demonstrate that the WLAMFormer model delivers state-of-the-art accuracy across various model scales, highlighting the effectiveness of integrating Discrete Wavelet Transform (DWT) into the Transformer architecture.

While achieving higher accuracy, the WLAMFormer model also maintains a competitive computational cost.

WLAMFormer_L1 exhibits a computational cost of 2.847 GFLOPs. While this is higher than that of some efficient models, such as EffiFormer-L1 [11] (1.310 GFLOPs) and SwiftFormer-L1 [45] (1.604 GFLOPs), it achieves a significant accuracy improvement of up to 3.8%.

 doi:10.20944/preprints202502.2130.v1

WLAMFormer_L2 provides an excellent balance with a computational cost of 3.803 GFLOPs. It outperforms ConvNeXt-T [48] by 2% in accuracy while reducing the FLOPs by 0.7G. Additionally, it surpasses Vmanba-T [49] by 1.6% in accuracy with a decrease of 1.1G in FLOPs, exceeds MLLA-T [25] by 0.6% in accuracy while reducing FLOPs by 0.4G, and outperforms WTConvNeXt-T [50] by 1.6% in accuracy with a reduction of 0.7G in FLOPs.

WLAMFormer_L3 achieves a high accuracy of 84.6% with a computational cost of 7.75 GFLOPs, exceeding ConvNeXt-S [48] by 1.5% in accuracy while reducing FLOPs by nearly 1G. It also surpasses Vmanba-S [49] by 1% in accuracy with an approximate 1G reduction in FLOPs, outperforms MLLA-S [25] by 0.2% in accuracy while decreasing FLOPs by 0.4G, and exceeds WTConvNeXt-T [50] by 1% in accuracy with a reduction of 1.1G in FLOPs.

The WLAMFormer model exhibits moderate performance in terms of throughput. **WLAMFormer_L1** achieves a throughput of 2296 images per second (imgs/s), surpassing CAS_ViT_M [46] (2254 imgs/s) but lagging behind other efficient models. The discrete wavelet transform and its inverse have an impact on image throughput, which is particularly noticeable in smaller models.

WLAMFormer_L2 delivers a throughput of 1580 imgs/s, exceeding that of Swin-T [3] (1246 imgs/s), CAS-ViT-T [46] (1084 imgs/s), and MLLA-T [25] (1009 imgs/s). However, it falls short compared to SwiftFormer-L1 [45] (5051 imgs/s) and EffiFormer-L1 [11] (5046 imgs/s), placing it at an intermediate level among models of comparable size.

WLAMFormer_L3 achieves a throughput of 881 imgs/s, which surpasses that of PVTv2-B3 [37] (403 imgs/s) and CSwin-S [41] (625 imgs/s), demonstrating a relatively strong performance among models of similar scale.

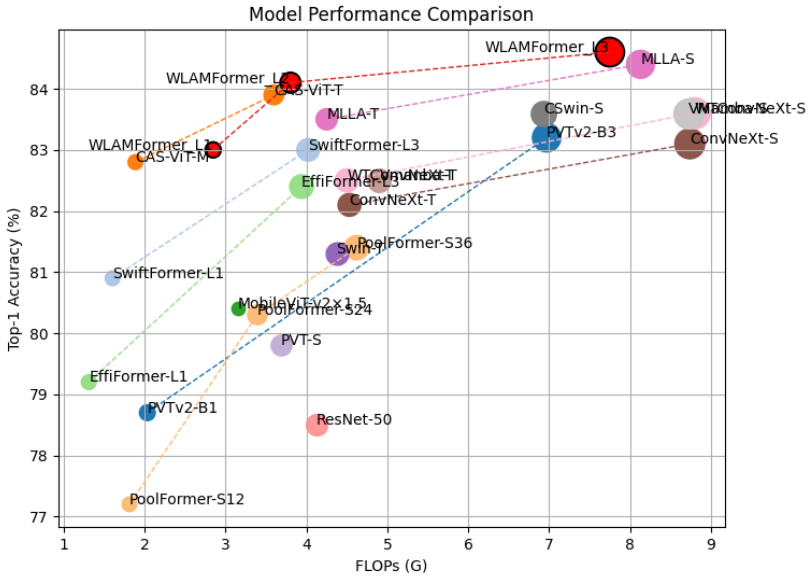


Figure 4. Comparison of Image Classification Performance on the ImageNet-1K Dataset.

The WLAMFormer model demonstrates a commendable balance between accuracy and computational efficiency. The integration of Discrete Wavelet Transform (DWT) enables the model to effectively capture multi-scale representations, achieving outstanding performance in image classification tasks. The slightly increased computational cost and reduced throughput represent a reasonable trade-off for the significant gains in accuracy, particularly in application scenarios where precision is paramount. However, in contexts with limited computational resources or stringent throughput requirements, the heightened computational demands and lower processing speeds may pose limitations. Future work could focus on optimizing the DWT operations and exploring more efficient implementation strategies to alleviate these drawbacks, thereby making the WLAMFormer model more suitable for a broader range of applications.

 doi:10.20944/preprints202502.2130.v1

In this paper, we introduce a plug-and-play attention module called WLAM (Wavelet-Enhanced Linear Attention Mechanism) and integrate it into mainstream neural network architectures, including PVT, Swin, and CSwin, to evaluate its performance enhancement in the ImageNet image classification task. Table 2 and Figure () provide a comparison of different models in terms of the number of parameters (Par.↓), FLOPs↓, and Top-1↑. In addition to the baseline models, we compare our approach with the plug-and-play Agent attention module proposed in [55], which has demonstrated exceptional performance in 2024. The results for all model variants are highlighted in bold.

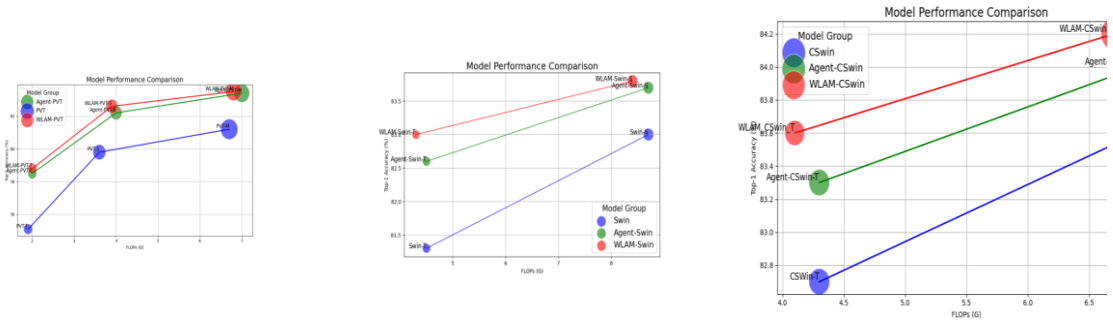
Table 2. Performance Comparison of Plug-and-Play Attention Modules.

Model	Par.↓(M)	Flops↓(G)	REs	Top-1↑
PVT-T	**11.2**	**1.9**	**224×224**	**75.1**
Agent-PVT-T	**11.6**	**2.0**	**224×224**	**78.5**
WLAM-**PVT-T**	**11.8**	**2.0**	**224×224**	**78.8**
PVT-S	**24.5**	**3.6**	**224×224**	**79.8**
Agent-PVT-S	**20.6**	**4.0**	**224×224**	**82.2**
WLAM-**PVT-S**	*20.8*	*3.9*	*224×224*	*82.6*
PVT-M	**44.2**	**6.7**	**224×224**	**81.2**
Agent-PVT-M	**35.9**	**7.0**	**224×224**	**83.4**
WLAM-**PVT-M**	*35.6*	*6.8*	*224×224*	*83.5*
Swin-T	**29**	**4.5**	**224×224**	**81.3**
Agent-Swin-T	**29**	**4.5**	**224×224**	**82.6**
WLAM-**Swin-T**	**27**	**4.3**	**224×224**	**83.0**
Swin-S	**50**	**8.7**	**224×224**	**83.0**
Agent-Swin-S	**50**	**8.7**	**224×224**	**83.7**
WLAM-**Swin-S**	**49**	**8.4**	**224×224**	**83.8**
CSWin-T	**23**	**4.3**	**224×224**	**82.7**
Agent-CSwin-T	**23**	**4.3**	**224×224**	**83.3**
WLAM_CSwin_T	**21**	**4.1**	**224×224**	**83.6**
CSwin-S	**35**	**6.9**	**224×224**	**83.6**
Agent-CSwin-S	**35**	**6.9**	**224×224**	**84.0**
WLAM-**CSwin-S**	**34**	**6.7**	**224×224**	**84.2**

- **Performance Improvement on the PVT Architecture：**
- Compared to the baseline model, WLAM-PVT-T exhibits a slight increase in parameters and FLOPs while achieving a 3.7 percentage point improvement in accuracy, surpassing Agent-PVT-T by 0.3 percentage points. This indicates that the WLAM module provides a more substantial performance enhancement in smaller models. WLAM-PVT-S, with parameters and FLOPs comparable to those of Agent-PVT-S, achieves an accuracy improvement of 0.4 percentage points over Agent-PVT-S and 2.8 percentage points over the baseline model, demonstrating the superiority of the WLAM module in mid-sized models. WLAM-PVT-M shows optimized parameters and FLOPs while achieving an accuracy that exceeds Agent-PVT-M by 0.1 percentage points and improves upon the baseline model by 2.3 percentage points, thereby validating the effectiveness of the WLAM module in large models.
- **Performance Improvement on the Swin Architecture**
- WLAM-Swin-T achieves a 1.7 percentage point increase in accuracy while reducing both parameters and computational load, outperforming the Agent version by 0.4 percentage points. This highlights the efficient performance of the WLAM module within the Swin-T model. WLAM-Swin-S demonstrates an accuracy increase of 0.8 percentage points over the baseline model and a 0.1 percentage point improvement compared to the Agent version, all while reducing parameters and FLOPs, further confirming the effectiveness of the WLAM module.
- **Performance Improvement on the CSwin Architecture**
- WLAM-CSwin-T achieves a 0.9 percentage point accuracy increase over the baseline model while reducing parameters and computational load, exceeding the Agent version by 0.3

 doi:10.20944/preprints202502.2130.v1

percentage points, which reflects the efficiency of the WLAM module. Similarly, WLAM-CSwin-S shows a 0.6 percentage point improvement in accuracy over the baseline model and a 0.2 percentage point increase compared to the Agent version, further showcasing the advantages of the WLAM module.



(a) (b) (c)

Figure 5. Performance Comparison of Plug-and-Play Attention Modules. (a) Comparison of Plug-and-Play Attention Performance Based on the PVT Model. (b) Comparison of Plug-and-Play Attention Performance Based on the Swin Model. (c) Comparison of Plug-and-Play Attention Performance Based on the CSwin Model.

- In both the PVT, Swin, and CSwin architectures, models integrated with the WLAM module achieved a significant improvement in Top-1 accuracy, with the maximum enhancement reaching 3.7 percentage points. Regarding parameter and computational efficiency, the WLAM models not only enhanced performance but also reduced the number of parameters and FLOPs in many cases, demonstrating their efficacy. Compared to models incorporating the Agent attention module, the WLAM models consistently achieved notable accuracy improvements, indicating that the WLAM module is superior in capturing feature representations.
- In addition to validating the performance of our network on ImageNet1K, we also tested our model on CIFAR-10[56] and CIFAR-100[56], both of which consist of low-resolution images, as illustrated in Table 3. We present a comparison of several publicly available models that report transfer accuracy on the CIFAR-10 and CIFAR-100 datasets. The parameters used for training our model on CIFAR-10 and CIFAR-100 are similar to those employed during training on ImageNet1K, specifically with 400 epochs and a batch size of 512, while maintaining other settings constant.

Table 3. This is a table. Tables should be placed in the main text near to the first time they are cited.

Model	Par.↓(M)	Flops↓(G)	Type	Top-1↑ (Cifar10)	Top-1↑ (Cifar100)
MobileViT-v2×1.5	**10.0**	**3.151**	**Hybrid**	**96.2**	79.5
EfficientFormer-L1[53]	12.3	2.4	**Hybrid**	97.5	83.2
EdgeViT-S[54]	11.1	1.1	**Transformer**	97.8	81.2
EdgeViT-M[54]	13.6	2.3	**Transformer**	98.2	82.7
PVT-Tiny	**11.2**	**1.9**	**Transformer**	**95.8**	**77.6**
WLAM-PVT-T	**11.8**	**2.0**	**DWT-Transformer**	**96.9**	**82.1**
WLAMFormer_L1	**13.5**	**2.8**	**DWT-Transformer**	**97.7**	**84.5**
PVT-Small	24.5	**3.8**	**Transformer**	**96.5**	**79.8**
WLAM-PVT-S	**20.8**	**3.9**	**DWT-Transformer**	**98.4**	**84.8**
PoolFormer-S24	21	3.5	**Pool**	96.8	81.8
EfficientFormer-L3[53]	31.9	5.3	**Hybrid**	98.2	85.7
ConvNeXt	28	**4.5**	**ConvNet**	**98.7**	**87.5**

 doi:10.20944/preprints202502.2130.v1

ConvNeXt V2-Tiny	28	**4.5**	**ConvNet**	99.0	**90.0**
EfficientNetV2-S	24	8.8	**ConvNet**	**98.1**	**90.3**
WLAMFormer_L2	**23**	**3.8**	**DWT-Transformer**	**98.2**	**87.1**

The WLAM-PVT-T model adds only a small number of parameters compared to the baseline PVT-T model (11.8M vs. 11.2M), yet it achieves a 4.5% improvement in accuracy on CIFAR-100 (from 77.6% to 82.1%) and a 1.2% increase on CIFAR-10 (from 95.8% to 97.7%). Similarly, the WLAM-PVT-S model incurs only a slight increase in FLOPs compared to the baseline PVT-S model (3.9G vs. 3.8G), while demonstrating a 5.0% enhancement in accuracy on CIFAR-100 (from 79.8% to 84.8%) and a 1.9% improvement on CIFAR-10 (from 96.5% to 98.4%). These results clearly indicate that the WLAM attention module significantly enhances the recognition capability of neural networks on low-resolution images.

The WLAMFormer_L1 achieves an accuracy of 84.5% on CIFAR-100, outperforming other models of similar scale, such as EfficientFormer-L1 (83.2%) and EdgeViT-M (82.7%). Due to the influence of wavelet transformations, the FLOPs value of WLAMFormer_L1 is relatively high among models of similar size (2.8G).

WLAMFormer_L2 reaches an accuracy of 98.2% on CIFAR-10 and 87.1% on CIFAR-100. Although its performance does not surpass that of ConvNet architectures such as ConvNeXt and EfficientNet, it demonstrates substantial improvements over non-CNN architectures, exceeding the accuracy of the PoolFormer-S24 model by 5.3% and the EfficientFormer-L3 model by 1.4%.

Traditional Transformer models (e.g., PVT-Tiny, PVT-Small) and hybrid models (e.g., EfficientFormer, EdgeViT) often underperform convolutional neural networks when processing small-sized images (such as those in the CIFAR dataset). This limitation arises because Transformer models require large amounts of data and higher resolutions to learn global features effectively. However, the WLAM series models introduce attention modules based on wavelet transformations, which effectively enhance the ability of Transformer models to capture multi-scale and multi-resolution features in small-sized images, facilitating the learning of critical detail information. The WLAM module applies linear attention to low-frequency components while employing convolutional.

4.2. Ablation Study

In this section, we investigate the effectiveness of key components within the WLAM attention module by systematically removing them. We report the results of ImageNet-1K classification based on the WLAMFormer_L2 model. ·

3. We removed the structure that mimics Mamba, while keeping all other components unchanged.
4. We discontinued the use of the structure that imitates MobileNetV3 for processing high-frequency subbands; instead, we employed a single 3x3 convolution for high-frequency subbands, similar to the approach outlined in [33].
5. We eliminated the multi-resolution input from the attention module, following the methodology of [33], and solely utilized the low-frequency components as inputs for linear attention. ·

$$Q = \phi(X_{LL}W_q), K = \phi(X_{LL}W_k), V = \phi\emptyset(X_{LL}W_v) \quad (29)$$

Table 4. Ablation Study Comparison.

Model	Par.↓(M)	Flops↓(G)	Top-1↑	difference
1	25.0	3.8	**82.9**	**-1.2**
2	**24.9**	4.2	83.3	-0.8
3	25.6	4.0	**82.1**	-2.0
WLAM**Former_L2**	**25.0**	**3.8**	84.1	—

Impact of the Mamba Biomimetic Structure (Model 1)

The removal of the Mamba-inspired structure led to a decrease in Top-1 accuracy from 84.1% to 82.9%, reflecting a reduction of 1.2%. This decline underscores the importance of the Mamba-inspired forget gate, which provides local bias and positional information to the attention module. The

 doi:10.20944/preprints202502.2130.v1

incorporation of learnable shortcuts in the Mamba design enhances the stability of the model. Its removal results in a significant performance drop, indicating that this component is critical for improving model accuracy.

Impact of the High-Frequency Processing Module (Model 2)

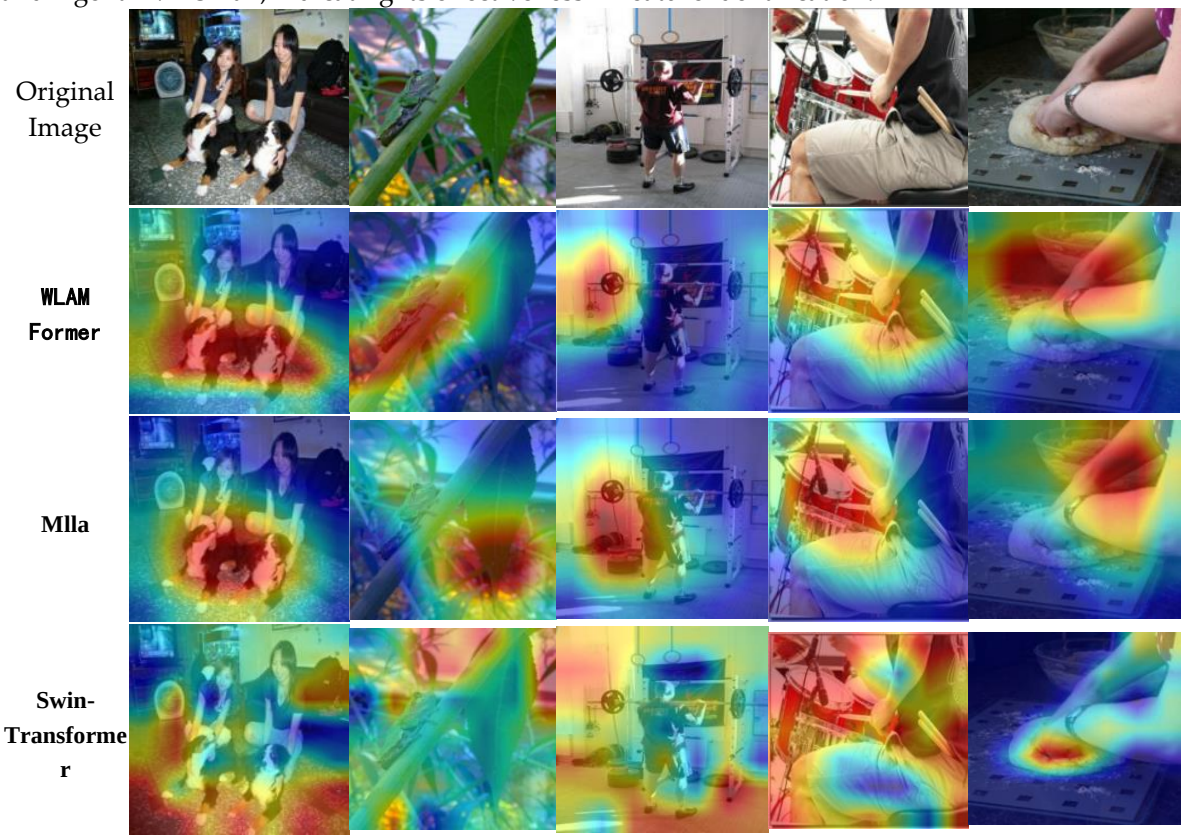
When the high-frequency processing module was simplified to a single 3×3 convolution, the Top-1 accuracy further declined to 83.3%, a reduction of 0.8%. The MobileNetV3-inspired high-frequency processing structure is designed to more effectively extract high-frequency detail features. Simplifying this module reduces the model's ability to capture fine-grained information, leading to a decrease in performance. However, this impact is comparatively less significant than that observed with the removal of the Mamba biomimetic structure.

Impact of Multi-Resolution Input in the Attention Module (Model 3)

The omission of the multi-resolution input, which limited the attention module to only low-frequency components, resulted in a Top-1 accuracy drop to 82.1%, a reduction of 2.0%. The multi-resolution input enables the attention module to integrate features across different scales, facilitating the fusion of global and local information. The removal of this component restricts the model's feature representation capabilities, leading to a substantial decline in performance. This effect represents the most significant impact observed across all ablation experiments.

4.3. Network Visualization

We employed the Grad-CAM method to generate heatmaps that highlight the regions of focus within the network. To validate the accuracy of the model's identification, we compared the heatmaps of WLAMFormer-L2, WLAM-Swin-T, and WLAM-PVT-Small with those of MLLA-T, Swin-T, Agent-Swin-T, PVT-Small, and Agent-PVT-Small. The results demonstrate that WLAMFormer-L2 exhibits a clear advantage in performance. Additionally, WLAM-Swin-T shows improved performance compared to both Swin-T and Agent-Swin-T. Similarly, WLAM-PVT-Small outperforms PVT-Small and Agent-PVT-Small, indicating its effectiveness in feature identification.



 doi:10.20944/preprints202502.2130.v1

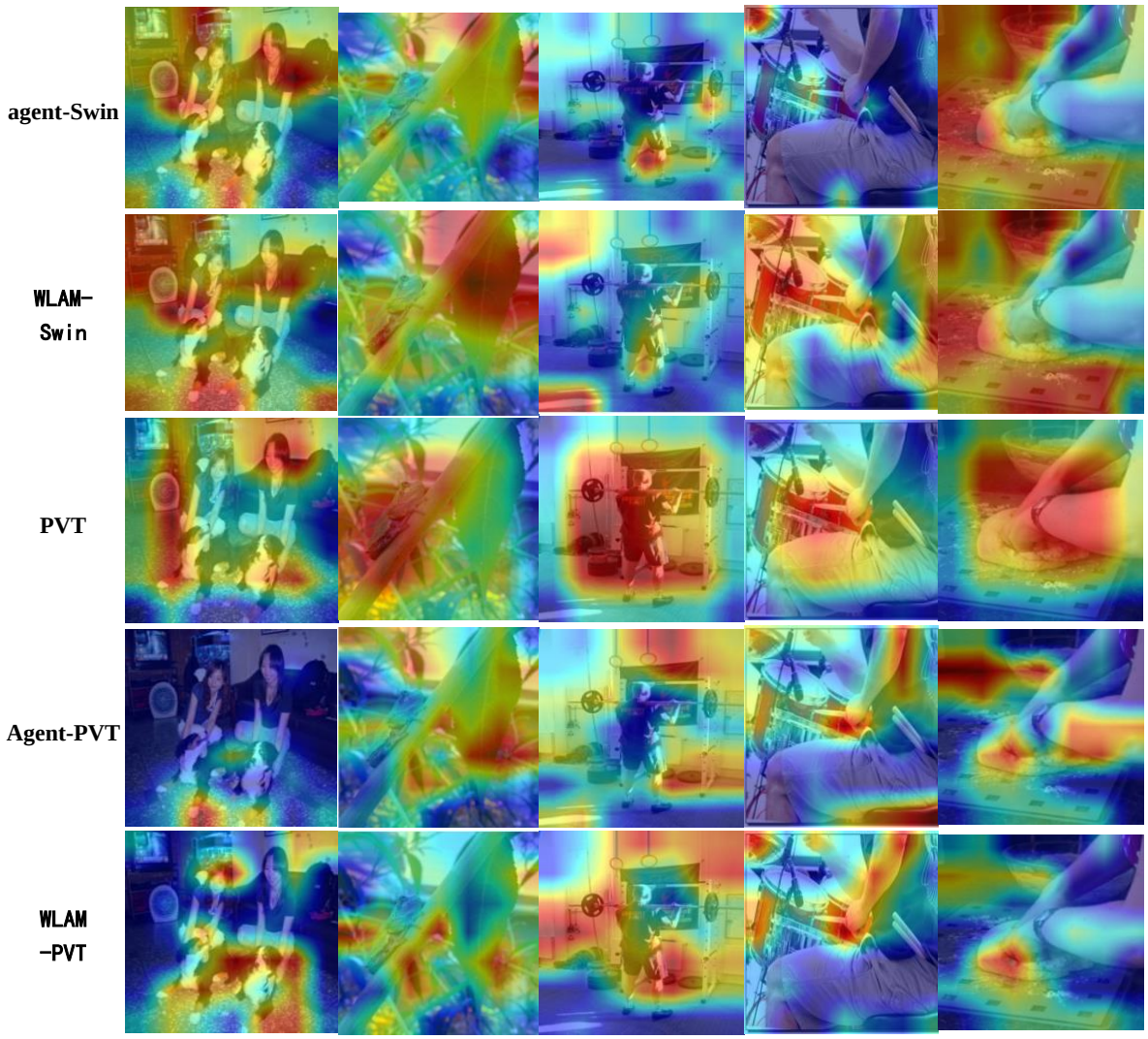


Figure 6. Heatmaps generated using the Grad-CAM method.

5. Conclusions

This paper addresses the limitations of linear attention in terms of performance by proposing a plug-and-play Wavelet-Enhanced Linear Attention Mechanism (WLAM) module. This module integrates Discrete Wavelet Transform (DWT) with linear attention to enhance the model's ability to express global context and local features. By introducing DWT into the attention mechanism, we perform wavelet decomposition on the input features, generating query vectors Q from the original input features, keys K from the low-frequency coefficients, and values V from the high-frequency coefficients processed through multi-scale convolutions and SE (Squeeze-and-Excitation) modules. This method effectively embeds global information and local features into different components of the attention mechanism, enhancing the model's perception of details and overall structure.

Furthermore, we reintegrate the multi-scale processed information back into the spatial domain using Inverse Discrete Wavelet Transform (IDWT), addressing the shortcomings of linear attention in handling multi-scale and local information. We also drew inspiration from the Mamba network's forget gate and improved block design, inheriting its core advantages to further enhance the model's performance and robustness. Based on the lossless downsampling characteristics of wavelet transforms, we proposed the Wavelet Downsampling Attention (WDSA) module, which combines downsampling and attention modules, reducing the network size and computational load while minimizing information loss due to downsampling. By combining the WLAM and WDSA modules, we constructed the WDLMFormer model. We applied the proposed WLAM module to classical networks such as PVT, Swin, and CSwin, significantly improving their performance on the image

 doi:10.20944/preprints202502.2130.v1

classification task of the ImageNet-1K dataset. The WDLMFormer achieved an accuracy of 84.6% on the ImageNet-1K dataset, validating the effectiveness and superiority of our approach.

In summary, the WLAM and WDSA modules proposed in this paper provide new insights into the design of attention mechanisms. By integrating wavelet transforms with linear attention, we successfully enhanced the model's capability to capture both global and local information, achieving outstanding performance in practical applications. However, there are still several issues worth further investigation and exploration. In future work, we will consider applying this method to more visual tasks, such as object detection and semantic segmentation, to validate its generality and effectiveness in different task scenarios. Additionally, we will explore more efficient ways to integrate wavelet transforms with deep learning models to further enhance model performance and computational efficiency.

Author Contributions: Conceptualization, B.F. and S.L.; methodology, B.F.; software, B.F.; validation, C.X., Z.L; formal analysis, B.F.; investigation, S.L.; resources, S.L.; data curation, S.L.; writing—original draft preparation, B.F.; writing—review and editing, B.F.; visualization, Z.L.; supervision, C.X.; project administration, C.X.; funding acquisition, C.X. All authors have read and agreed to the published version of the manuscript.

Funding: National Key Research and Development Program funded project. 2019YFC0117800.

Data Availability Statement: The ImageNet1K dataset can be downloaded from the website https://www.image-net.org/. The CIFAR-10 and CIFAR-100 datasets can be downloaded from the website https://www.cs.toronto.edu/~kriz/cifar.html.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table 5. Structure of the WLAM_Swin and WLAM_CSwin.

Stage	Output	WLAM_Swin_T		WLAM_Swin_S		WLAM_Swin_B	
		WLAM_Block	**Swin_Block**	WLAM_Block	**Swin_Blocck**	WLAM_Block	**Swin_Blocck**
1	**56×56**	**Concat 4×4,96,LN**		**Concat 4×4,96,LN**		**Concat 4×4,128,LN**	
		[win 56 × 56 dim 96 head 3]	[win 56 × 56 dim 96 head 3]	[win 56 × 56 dim 96 head 3]	[win 56 × 56 dim 96 head 3]	[win 56 × 56 dim 128 head 3]	[win 56 × 56 dim 128 head 3]
2	**28×28**	**Concat 4×4,192,LN**		**Concat 4×4,192,LN**		**Concat 4×4,256,LN**	
		[win 28 × 28 dim 192 head 6]	[win 28 × 28 dim 192 head 6]	[win 28 × 28 dim 192 head 6]	[win 28 × 28 dim 192 head 6]	[win 28 × 28 dim 256 head 6]	[win 28 × 28 dim 256 head 6]
3	**14×14**	**Concat 4×4,384,LN**		**Concat 4×4,384,LN**		**Concat 4×4,512,LN**	
		[win 14 × 14 dim 384 head 12] **×3**	[win 14 × 14 dim 384 head 12] **×3**	[win 14 × 14 dim 384 head 12] **×9**	[win 14 × 14 dim 384 head 12] **×9**	[win 14 × 14 dim 512 head 12] **×9**	[win 14 × 14 dim 512 head 12] **×9**
4	**7×7**	**Concat 4×4,768,LN**		**Concat 4×4,768,LN**		**Concat 4×4,1024,LN**	
		None	[win 7 × 7 dim 768 head 24] **×2**	**None**	[win 7 × 7 dim 768 head 24] **×2**	**None**	[win 7 × 7 dim 1024 head 24] **×2**
Stage	**Output**	WLAM_CSwin_T		WLAM_CSwin_S		WLAM_CSwin_B	
		WLAM_Block	**CSwin_Blocck**	WLAM_Block	**CSwin_Blocck**	WLAM_Block	**CSwin_Blocck**

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 February 2025 doi:10.20944/preprints202502.2130.v1

1	56×56	Concat 7×7,stride=4,64,LN		Concat 7×7,stride=4,64,LN		Concat 7×7,stride=4,96,LN	
		[win 56 × 56, dim 64, head 2]	[win56 × 56, dim 64, head 2]	[win 56 × 56, dim 64, head 2]	[win 56 × 56, dim 64, head 2] × 2	[win 56 × 56, dim 96, head 2]	[win 56 × 56, dim 96, head 2] × 2
2	28×28	Concat 7×7,stride=4,128,LN		Concat 7×7,stride=4,128,LN		Concat 7×7,stride=4,192,LN	
		[win 28 × 28, dim 128, head 4] × 2	[win 28 × 28, dim 128, head 4] × 2	[win 28 × 28, dim 128, head 4] × 3	[win 28 × 28, dim 128, head 4] × 3	[win 28 × 28, dim 192, head 4] × 3	[win 28 × 28, dim 192, head 4] × 3
3	14×14	Concat 7×7,stride=4,256,LN		Concat 7×7,stride=4,256,LN		Concat 7×7,stride=4,384,LN	
		[win 14 × 14, dim 256, head 8] ×9	[win 14 × 14, dim 256, head 8] ×9	[win 14 × 14, dim 256, head 8] ×15	[win 14 × 14, dim 256, head 8] ×14	[win 14 × 14, dim 384, head 8] ×15	[win 14 × 14, dim 384, head 8] ×14
4	7×7	Concat 7×7,stride=4,512,LN		Concat 7×7,stride=4,512,LN		Concat 7×7,stride=4,768,LN	
		None	[win 7 × 7, dim 512, head 16] ×1	None	[win 7 × 7, dim 512, head 16] ×2	None	[win 7 × 7, dim 768, head 16] ×2

Appendix B

Table 6. Structure of the MLAMFormer Model.

Stage	Output	WLAMFormer_L1		WLAMFormer_L2		WLAMFormer_L3	
		WLAM_Block	Liner_Block	WLAM_Block	Liner_Block	WLAM_Block	Liner_Block
1	56×56	stem,32		stem,32		stem,42	
		Attention_DownSampling,64		Attention_DownSampling,64		Attention_DownSampling,84	
		[win 56 × 56, dim 64, head 2]	None	[win 56 × 56, dim 64, head 2]	None	[win 56 × 56, dim 84, head 3]	None
2	28×28	Attention_DownSampling,128		Attention_DownSampling,128		Attention_DownSampling,168	
		[win 28 × 28, dim 128, head 4]	None	[win 28 × 28, dim 128, head 4] × 3	None	[win 28 × 28, dim 168, head 6] × 3	None
3	14×14	Attention_DownSampling,256		Attention_DownSampling,256		Attention_DownSampling,336	
		[win 14 × 14, dim 256, head 8] ×3	[win 14 × 14, dim 256, head 12] ×2	[win 14 × 14, dim 256, head 8] ×4	[win 14 × 14, dim 256, head 8] ×3	[win 14 × 14, dim 226, head 12] ×6	[win 14 × 14, dim 226, head 12] ×5

 doi:10.20944/preprints202502.2130.v1

4	**7×7**	Attention_**DownSampling,512**	Attention_**DownSampling,512**	Attention_**DownSampling,672**
		None $\begin{bmatrix}\text{win } 7\times 7\\ \text{dim } 512\\ \text{head } 16\end{bmatrix}$ **×2**	**None** $\begin{bmatrix}\text{win } 7\times 7\\ \text{dim } 512\\ \text{head } 16\end{bmatrix}$ **×3**	**None** $\begin{bmatrix}\text{win } 7\times 7\\ \text{dim } 672\\ \text{head } 24\end{bmatrix}$ **×3**

References

1. Alexey, Dosovitskiy. "An image is worth 16x16 words: Transformers for image recognition at scale." arxiv preprint arxiv: 2010.11929 (2020).
2. Zhao, S.; Yang, J.; Wu, N.; Wu, Y.;& Zhang, T. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions.Proceedings of the IEEE/CVF international conference on computer vision, 2021: 568-578.
3. Ze, L; Yutong, L;Yue,C; Han,H; Yixuan, W; Zheng,Z; Stephen, L; Baining,G.Swin transformer: Hierarchical vision transformer using shifted windows.In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021;pp. 10012-10022.[CrossRef]
4. Liu, Z.;Lin,Y.;Cao,Y.;Hu,H.;Wei,Y.;Zhang,Z.;&Guo,B. Swin transformer v2: Scaling up capacity and resolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022;pp. 12009-12019.[CrossRef]
5. Xia,Z.; Pan, X.;Song, S.; Li, L. E.; & Huang, G. . Vision transformer with deformable attention. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,2022;pp. 4794-4803.
6. Zhou, H.; Zhang, Y.; Guo, H.; Liu, C.; Zhang, X.; Xu, J.;& Gu, J.. Neural architecture transformer. arXiv preprint arXiv:2106.04247(2021).
7. Zhu, L.; Wang, X.; Ke, Z.;Zhang, W.;&Lau, R. W. . Biformer: Vision transformer with bi-level routing attention. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,2023;pp. 10323-10333.[CrossRef]
8. Wang, S.;Li, B.Z.;Khabsa, M.;Fang, H.;&Ma,H.. Linformer: Self-attention with linear complexity. arxiv preprint arxiv:2006.04768(2020).[CrossRef]
9. Qin, Z.; Sun, W.; Deng, H.; Li, D.; Wei, Y.; Lv, B.; Zhong, Y.. cosformer: Rethinking softmax in attention. arxiv preprint arxiv:2202.08791 (2022).[CrossRef]
10. Ma, X.; Kong, X.; Wang, S.; Zhou, C.; May, J.; Ma, H.;& Zettlemoyer, L. . Luna: Linear unified nested attention. Advances in Neural Information Processing Systems,2021; 34, 2441-2453.
11. Shen, Z.; Zhang, M.; Zhao, H.; Yi, S.; & Li, H. . Efficient attention: Attention with linear complexities. In Proceedings of the IEEE/CVF winter conference on applications of computer vision,2021;pp. 3531-3539.[CrossRef]
12. Gao, Y.; Chen, Y.; & Wang, K. . SOFT: A simple and efficient attention mechanism. arXiv preprint arXiv:2104.02544(2021).
13. Xiong, Y.; Zeng, Z.; Chakraborty, R.; Tan, M.; Fung, G.; Li, Y.; & Singh, V.. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*,2021; May,Vol. 35, No. 16, pp. 14138-14148.[CrossRef]
14. Haoran, You; Yunyang, Xiong; Xiaoliang, Dai; Bichen, Wu; Peizhao, Zhang; Haoqi, Fan; Peter, Vajda; Yingyan, (Celine) Lin . Castling-vit: Compressing self-attention via switching towards linear-angular attention at vision transformer inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition ,2023;pp. 14431-14442.
15. Han, D.; Pan, X.; Han, Y.; Song, S.; & Huang, G.. Flatten transformer: Vision transformer using focused linear attention. In *Proceedings of the IEEE/CVF international conference on computer vision* ,2023;pp. 5961-5971.[CrossRef]
16. Han, D.; Ye, T.; Han, Y.; Xia, Z.; Pan, S.; Wan, P.;& Huang, G.. Agent attention: On the integration of softmax and linear attention. In *European Conference on Computer Vision* ,2024; September,pp. 124-140. Cham: Springer Nature Switzerland.

 doi:10.20944/preprints202502.2130.v1

17. Xu, Z.; Wu, D.; Yu, C.; Chu, X.; Sang, N.; & Gao, C.. SCTNet: Single-Branch CNN with Transformer Semantic Information for Real-Time Segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence,2024; March, Vol. 38, No. 6, pp. 6378-6386.[CrossRef]
18. Jiang, J.; Zhang, P.; Luo, Y.; Li, C.; Kim, J. B.;Zhang, K.; Kim, S.. AdaMCT: adaptive mixture of CNN-transformer for sequential recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* ,2023; October,pp. 976-986.[CrossRef]
19. Lou, M.; Zhou, H. Y.; Yang, S.; & Yu, Y. . TransXNet: learning both global and local dynamics with a dual dynamic token mixer for visual recognition. *arxiv preprint arxiv:2310.19380*(2023). [CrossRef]
20. Yoo, J.; Kim, T.; Lee, S.; Kim, S. H.; Lee, H.; & Kim, T. H.. Enriched cnn-transformer feature aggregation networks for super-resolution. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*,2023;pp. 4956-4965. [CrossRef]
21. Maaz, M.; Shaker, A.; Cholakkal, H.; Khan, S.; Zamir, S. W.; Anwer, R. M.; & Shahbaz Khan, F.. Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *European conference on computer vision*,2022; October,pp. 3-20. Cham: Springer Nature Switzerland.
22. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; & Zhang, L.. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* ,2021;pp. 22-31.[CrossRef]
23. Mehta, S.; Rastegari, M. . Separable self-attention for mobile vision transformers. *arxiv preprint arxiv:2206.02680*(2022).[CrossRef]
24. Wadekar, S. N.; & Chaurasia, A. . MobileViTv3: Mobile-friendly vision transformer with simple and effective fusion of local, global, and input features. *arXiv preprint arXiv:2209.15159*(2022).[CrossRef]
25. Han, D.; Wang, Z.; Xia, Z.; Han, Y.; Pu, Y.; Ge, C.;& Huang, G..Demystifying Mamba in Vision: A Linear Attention Perspective.arXiv:2405.16605, 2024.[CrossRef]
26. Bae, W.; Yoo, J.; & Chul Ye, J.. Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*,2017;pp. 145-153.[CrossRef]
27. Fujieda, S.;Takayama, K.; Hachisuka, T..Wavelet convolutional neural networks. arXiv preprint arXiv:1805.08620 (2018). [CrossRef]
28. Yao, T.; Pan, Y.; Li, Y.; Ngo, C. W.; & Mei, T. . Wave-vit: Unifying wavelet and transformers for visual representation learning. In *European Conference on Computer Vision*,2022;October,pp. 328-345. Cham: Springer Nature Switzerland.
29. Li, J.; Cheng, B.; Chen, Y.;Gao, G.; Shi, J.; & Zeng, T.. EWT: Efficient Wavelet-Transformer for single image denoising. *Neural Networks,*2024; *177*, 106378.
30. Azad, R.; Kazerouni, A.; Sulaiman, A.; Bozorgpour, A.; Aghdam, E. K.; Jose, A.; & Merhof, D.. Unlocking fine-grained details with wavelet-based high-frequency enhancement in transformers. In *International Workshop on Machine Learning in Medical Imaging* ,2023;October,pp. 207-216. Cham: Springer Nature Switzerland.
31. Gao, X.; Qiu, T.; Zhang, X.; Bai, H.; Liu, K.; Huang, X.; Liu, H.. Efficient multi-scale network with learnable discrete wavelet transform for blind motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,2024;pp. 2733-2742.[CrossRef]
32. Roy, A.; Sarkar, S.; Ghosal, S.; Kaplun, D.; Lyanova, A.; & Sarkar, R.. A wavelet guided attention module for skin cancer classification with gradient-based feature fusion. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)* ,2024; May,pp. 1-4. IEEE.
33. Koonce, B.; Koonce, B. E. *Convolutional neural networks with swift for tensorflow: Image recognition and dataset categorization* ,2021.;pp. 109-123. New York, NY, USA: Apress.
34. Han, D.; Wang, Z.; Xia, Z.; Han, Y.; Pu, Y.; Ge, C.; Huang, G.. Demystify Mamba in Vision: A Linear Attention Perspective. *arxiv preprint arxiv:2405.16605* (2024).[CrossRef]
35. Finder, S. E.; Amoyal, R.; Treister, E.; & Freifeld, O. . Wavelet convolutions for large receptive fields. In *European Conference on Computer Vision*,2024; September,pp. 363-380. Cham: Springer Nature Switzerland.

 doi:10.20944/preprints202502.2130.v1

36. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.;& Wang, X. . Vision mamba: Efficient visual representation learning with bidirectional state space model. *arxiv preprint arxiv:2401.09417*(2024).[CrossRef]
37. Wang, W.; Xie, E.; Li, X.; Fan, D. P.; Song, K.;Liang, D.; Shao, L.. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media* ,2022; *8*(3), 415-424.
38. Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Guo, B.. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* ,2022;pp. 12009-12019.[CrossRef]
39. Tan, J.; Pei, S.; Qin, W.; Fu, B.; Li, X.; & Huang, L.. Wavelet-based Mamba with Fourier Adjustment for Low-light Image Enhancement. In *Proceedings of the Asian Conference on Computer Vision* ,2024;pp. 3449-3464.[CrossRef]
40. Deng, J.; Dong, W.; Socher, R.; Li, L. J.; Li, K.; & Fei-Fei, L. . Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*,2009; June,pp. 248-255. Ieee.
41. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Guo, B.. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,2022;pp. 12124-12134.[CrossRef]
42. ——, "Sgdr: Stochastic gradient descent with warm restarts," in International Conference on Learning Representations, 2016.
43. Szegedy, C.; Vanhoucke, V.; Ioffe, S., Shlens, J.; & Wojna, Z.. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* ,2016;pp. 2818-2826.
44. Cubuk, E. D.; Zoph, B.; Shlens, J.; Le, Q. V.. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020 ; pp. 702-703.
45. Shaker, A.; Maaz, M.; Rasheed, H.; Khan, S.; Yang, M. H.; & Khan, F. S.. Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* ,2023;pp. 17425-17436.[CrossRef]
46. Zhang, T.; Li, L.; Zhou, Y.; Liu, W.; Qian, C.; & Ji, X.. Cas-vit: Convolutional additive self-attention vision transformers for efficient mobile applications. *arxiv preprint arxiv:2408.03703*(2024).[CrossRef]
47. Yu, W.; Luo, M.; Zhou, P.;Si, C.;Zhou, Y.;Wang, X.;Yan, S. . Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* ,2022; pp. 10819-10829.[CrossRef]
48. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I. S.. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* ,2023;pp. 16133-16142.
49. Huang, T.; Pei, X.; You, S.; Wang, F.; Qian, C.; & Xu, C.. Localmamba: Visual state space model with windowed selective scan. *arxiv preprint arxiv:2403.09338*(2024).[CrossRef]
50. Finder, S. E.; Amoyal, R.; Treister, E.; & Freifeld, O.. Wavelet convolutions for large receptive fields. In *European Conference on Computer Vision*,2024; September,pp. 363-380. Cham: Springer Nature Switzerland.
51. Han, D.;Ye, T.; Han, Y.; Xia, Z.; Pan, S.; Wan, P.; Huang, G.. Agent attention: On the integration of softmax and linear attention. In European Conference on Computer Vision ,2024; September,pp. 124-140. Cham: Springer Nature Switzerland.
52. Krizhevsky, A.; Hinton, D. . Learning multiple layers of features from tiny images. Technical Report, University of Toronto(2009).
53. X. Liu; H. Peng; N. Zheng; Y. Yang; H. Hu; and Y. Yuan.. Efficientvit:Memory efficient vision transformer with cascaded group attention.In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023; pp. 14 420–14 430.
54. J. Pan; A. Bulat; F. Tan; X. Zhu; L. Dudziak; H. Li;G. Tzimiropoulos; and B. Martinez.. Edgevits: Competing light-weight cnns on mobile devices with vision transformers. In European Conference on Computer Vision. Springer, 2022; pp. 294–311.

 doi:10.20944/preprints202502.2130.v1